Kosmos-2: Grounding Multimodal Large Language Models to the World

Year: 2023 | *Citations: 992* | *Authors: Zhiliang Peng, Wenhui Wang, Li Dong*

Abstract

We introduce Kosmos-2, a Multimodal Large Language Model (MLLM), enabling new capabilities of perceiving object descriptions (e.g., bounding boxes) and grounding text to the visual world. Specifically, we represent refer expressions as links in Markdown, i.e., ``[text span](bounding boxes)'', where object descriptions are sequences of location tokens. Together with multimodal corpora, we construct large-scale data of grounded image-text pairs (called GrIT) to train the model. In addition to the existing capabilities of MLLMs (e.g., perceiving general modalities, following instructions, and performing in-context learning), Kosmos-2 integrates the grounding capability into downstream applications. We evaluate Kosmos-2 on a wide range of tasks, including (i) multimodal grounding, such as referring expression comprehension, and phrase grounding, (ii) multimodal referring, such as referring expression generation, (iii) perception-language tasks, and (iv) language understanding and generation. This work lays out the foundation for the development of Embodiment AI and sheds light on the big convergence of language, multimodal perception, action, and world modeling, which is a key step toward artificial general intelligence. Code and pretrained models are available at https://aka.ms/kosmos-2.